Test-retest reliability of the Isernhagen Work Systems Functional Capacity Evaluation in patients with chronic back pain

Brouwer, S.; Reneman, M.F.; Dijkstra, P.U.; Groothoff, J.W.; Schellekens, J.M.H.; Göeken, L.N.H.

Published in:
Journal of Occupational Rehabilitation

Publication date:
2003

Journal of Occupational Rehabilitation, Vol. 13, No. 4, December 2003 (© 2003)

Test–Retest Reliability of the Isernhagen Work Systems Functional Capacity Evaluation in Patients With Chronic Low Back Pain

S. Brouwer,[1,6] **M. F. Reneman,**[1] **P. U. Dijkstra,**[1,2] **J. W. Groothoff,**[3] **J. M. H. Schellekens,**[4] **and L. N. H. Göeken**[1,5]

The aim of this study was to investigate test–retest reliability of the Isernhagen Work System Functional Capacity Evaluation (IWS FCE) in a sample of patients ($n = 30$) suffering from Chronic Low Back Pain (CLBP) and selected for rehabilitation treatment. The IWS FCE consists of 28 tests that reflect work-related activities like lifting, carrying, bending, etc. In this study, a slightly modified IWS FCE was used. Patients were included in the study if they were still at work or were less than 1 year out of work because of CLBP. Participants' mean age was 40 years, the duration of low back pain ranged between 5 and 10 years. Fifteen patients (50%) were out of work for a mean of 17 weeks, and they all received financial compensation. Two FCE sessions were held with a 2-week interval in between. Means per session, 95% confidence intervals of the mean difference, one-way random Intra Class Correlations (ICC), limits of agreement, Cohen's kappa and percentage of absolute agreement were calculated where appropriate. An ICC of 0.75 or more, a kappa value of more than 0.60 and a percentage of absolute agreement of 80% were considered as an acceptable reliability. Tests of the IWC FCE were divided into tests with and tests without an acceptable test–retest reliability on the basis of the kappa values, the percentage of absolute agreement and the ICC values. Fifteen tests (79%) showed an acceptable test–retest reliability based on Kappa values and percentage of absolute agreement. Eleven tests (61%) showed an acceptable test–retest reliability based on ICC values.

KEY WORDS: reliability; chronic low back pain; functional capacity evaluation; disability assessment; work.

INTRODUCTION

Chronic low back pain (CLBP) is an endemic problem in modern Western society. The costs of chronic low back pain are enormous and can be attributed to direct and indirect costs

[1]Center for Rehabilitation, Groningen University Hospital, The Netherlands.
[2]Department of Oral and Maxillofacial Surgery, Groningen University Hospital, The Netherlands.
[3]Northern Center for Health Care Research, University of Groningen, The Netherlands.
[4]Department of Experimental and Occupational Psychology, University of Groningen, The Netherlands.
[5]Institute for Movement Sciences, University of Groningen, The Netherlands.
[6]Correspondence should be directed to Sandra Brouwer, Center for Rehabilitation, Groningen University Hospital, P.O. Box 30.001, 9700 RB Groningen, The Netherlands; e-mail: s.brouwer@rev.azg.nl.

1053-0487/03/1200-0207/0 © 2003 Plenum Publishing Corporation

of health care. Many patients suffering from CLBP are unable to work, thereby contributing to the total costs associated with CLBP. A major aim of rehabilitation treatment in these patients is to improve work ability. Therefore, in order to evaluate treatment outcome, the ability to work has to be assessed. Assessment of work-related abilities and disabilities in patients suffering from CLBP is not an easy task. To assess an individual's functional abilities to work, a number of functional capacity evaluations (FCEs) are available. FCE's are test batteries aimed at measuring functional abilities to work safely and productively (1).

One of the more well-known FCEs is the Isernhagen Work System (IWS) FCE. The IWS FCE consists of 28 tests that reflect work-related activities such as lifting, carrying, bending, etc. These tests are based on the job factors of the Dictionary of Occupational Titles (DOT), a publication of the United States Department of Labor (1,2). This dictionary describes the physical activities (job factors) that a job requires in a systematic way, by means of physical demands analysis.

To determine whether the IWS FCE can be used as an instrument to assess work-related rehabilitation outcome in patients suffering from CLBP, the reliability of the instrument, amongst other psychometric properties, has to be known. Parts of the IWS FCE already have been tested for their reliability. In a test–retest design, "lifting" and "carrying" have been found to possess a good reliability (3,4). The intraclass correlations ranged from 0.77 to 0.94. Pushing static and pulling static also appeared to possess good test–retest reliability (5), as does the measurement of maximum holding times (6). However, no studies are available that investigated all tests of the IWS FCE. The aim of this study was to investigate test–retest reliability of all tests of the IWS FCE in a sample of patients suffering from CLBP.

METHODS

Subjects

Subjects were 30 consecutive patients diagnosed with CLBP (24 males, 6 females), selected for rehabilitation treatment by physiatrists of the Center for Rehabilitation at Groningen University Hospital and who agreed to participate. Inclusion criteria were non-specific CLBP and being still at work or less than 1 year out of work because of CLBP. All patients were referred for treatment in a rehabilitation center between May 2000 and April 2001. The mean age of the patients was 40 (SD = 8.1 years). The duration of low back pain ranged between 5 and 10 years. Fifteen patients (50%) were out of work and all of them were receiving financial compensation. Patients were out of work for a mean of 17 weeks (SD = 19.2).

Procedure

Demographics and medical history were obtained of all subjects. Two FCE sessions were held with a 2-week interval. After an introduction of the FCE procedures and after signing informed consent, the patients were briefly instructed on how to perform each test. The evaluator first demonstrated each test. In this way, a total of 28 tests were performed (Table I). The patients were asked to perform the tests to the maximum of their abilities. Testing could be terminated for four reasons: 1) It was explained that they were allowed to stop the procedures at any point if they wished to do so, for example because of insecurity or pain; 2) A heart rate monitor was worn by the patients throughout the test procedures. A test

Table I. Description of the activities of the Isernhagen Work Systems (IWS) FCE

FCE activity	Description	Scoring
Lifting	5 lifts from table to floor v.v.; 4–5 weight increments; <90 s.	Max amount kg lifted
Overhead lift	5 lifts from table to crown height v.v.; 4–5 weight increments; <90 s.	Max amount kg lifted
Short carry two-handed	5 carries 1.5 m.; waist height; 4–5 weight increments; <90 s.	Max amount kg carried
Long carry two-handed	1 carry 20 m.; waist height; 4–5 weight increments; <90 s.	Max amount kg carried
Long carry right-handed	1 carry 20 m.; waist height; 4–5 weight increments; <90 s.	Max amount kg carried
Long carry left-handed	1 carry 20 m.; waist height; 4–5 weight increments; <90 s.	Max amount kg carried
Pushing static	Static full body push; 3 repetitions	average kgF
Pulling static	Static full body pull; 3 repetitions	average kgF
Pushing dynamic	Pushing a weighted cart over 10 m including 2 turns	Safely possible yes/no
Pulling dynamic	Pulling a weighted cart over 10 m including 2 turns	Safely possible yes/no
Overhead work test*	Standing with hands at crown height; manipulating nut/bolts	Time position is held (s)
Forward bend test standing*	Standing with 30–60° trunk flexion; manipulating nut/bolts	Time position is held (s)
Forward bend test sitting	Sitting with 30-60° trunk flexion; manipulating nut/bolts, max. 5 min	Time position is held (s)
Kneeling	Maintaining kneeling posture; knees 90° flexion, hips straight, max. 5 min.	Time position is held (s)
Crawling	Ambulate 3 m on hands and knees, then replace small object from floor to table height while in crawling position; 10 reps.	Able yes/no
Crouching	Maintaining position with knees and hips fully flexed, max 1 min.	Time position is held (s)
Dynamic bending*	Repetitive bending at hips and back; remove small object from floor to crown height 20 reps.	Time needed to perform 20 reps (s)
Dynamic squatting	Repetitive squatting with full flexion at knees and hips; remove small object from floor to crown height 20 reps.	Time needed to perform 20 reps (s)
Rep. rotation standing right*	Remove object horizontally at table height from left to right with left hand/arm; distance wing span; 30 reps.; standing.	Time needed to perform 30 reps (s)
Rep. rotation standing left*	Remove object horizontally at table height from right to left with right hand/arm; distance: wing span; 30 reps.; standing.	Time needed to perform 30 reps (s)
Rep. rotation sitting right*	Remove object horizontally at table height from left to right with left hand/arm; distance wing span; 30 reps.; sitting.	Time needed to perform 30 reps (s)
Rep. rotation sitting left*	Remove object horizontally at table height from right to left with right hand/arm; distance: wing span; 30 reps.; sitting.	Time needed to perform 30 reps (s)
Walking*	Shuttle walk test; increase speed per minute	Max. meters walked
Stair-climbing	Ascend and descent 100 steps; no handrail;	Able yes/no
Ladder-climbing*	Ascend and descent stepladder with 5 steps with use of hands	Able yes/no
Balance	Walking over a 10 × 300 cm balance board; forward, backward, heel to toe, sideways (6 ways; total mistakes)	Able with less than 6 mistakes yes/no
Sitting tolerance	30 min uninterrupted sitting, minor weight shifts allowed	Able yes/no
Standing tolerance	30 min uninterrupted standing, minor weight shifts allowed	Able yes/no

Note. (Items marked (*) are modified from the standard IWS FCE protocol) Rep: repetitive; max: maximal; sec: seconds; kg: kilograms; kgF: kilograms force; v.v.: vice versa; m: meters.

was terminated when the patient's heart rate met or exceeded 85% of his/her age-related maximum; 3) The evaluator terminated testing if it became unsafe. Unsafety was defined as a situation in which the patient was not in full control of himself/herself and/or the load; 4) For some tests, a time limit caused the patient to stop (i.e. crouching, max 60 s). After each test procedure, the evaluator recorded the results. One evaluator who had completed a formal FCE training course and had performed over 200 FCEs evaluated all FCEs. Time, day, and place of assessment were held constant for the two FCE sessions. Each session lasted between 2–3 h. The present study was approved by the institutional review board.

In this study, a modified IWS FCE was used (7). Instead of two consecutive days of testing, one test session was performed on a single day, because the test results on the second day only marginally differed from those of the first day (8). The tests on pushing and pulling dynamic, crawling, walking, stair climbing, and ladder climbing were slightly modified (Table I). Minor modifications were made to the overhead work test and the forward bend test standing, with patients being instructed to hold these postures as long as possible (6). The ceiling of these tests was set at 15 instead of 5 min because otherwise too many patients reached this ceiling and did not perform to the maximum of their capacities. Other tests that were modified are marked (*) in Table I. The performance criteria were stricter than the original protocols, without changing the essence of the tests. The scoring of the tests was elevated from a dichotomous (able yes/no) to an interval level (seconds needed to perform an item). This enabled us to use more powerful statistics. Grip strength and hand coordination tests were excluded because it was assumed that these tests could not be limited due to CLBP.

Data Analysis

Three types of tests can be distinguished in the IWS FCE: those without criterion or ceiling, those with a criterion, and those with a ceiling. Tests without criterion or ceiling are the material-handling tests and the shuttle-walk test. For the material-handling tests and shuttle-walk test, means, standard deviations, 95% confidence intervals, one-way random Intra Class Correlations (ICC), and limits of agreement were calculated (9).

A criterion indicates that a test is fulfilled when a criterion is met. For instance, the test "pushing dynamic" has a criterion that a subject is able or unable to safely push a weighted cart over a distance of 20 m. Repetitive rotation also has a criterion—the time needed to perform 30 rotations.

A ceiling indicates that the test is terminated because the patient had met what is defined to be the maximal time of performance. For instance, the overhead work test has a ceiling at 15 min. The test was terminated when the subject reached 15 min. In that case, patients have not performed to their maximal ability.

For all tests with criterion or ceiling effect, the number of subjects was determined who met the criterion or the ceiling for each test session. On the basis of these dichotomous results, Cohen's kappa's were calculated as well as percentages of absolute agreement of subjects with identical test behavior. Cohen's kappa's could not be calculated when the filling of the 2×2 tables was incomplete.

For the tests with a criterion, the following additional procedure was followed: if a patient reached the criterion in sessions 1 and 2, the times needed to reach the criterion in the sessions were used for further analyses, and means, standard deviations, 95% confidence

intervals, ICCs, and limits of agreement were calculated. Data of patients not meeting the criterion were excluded from the analyses. No further analyses were performed on dichotomous data.

For tests with a ceiling, another procedure was followed. If a patient reached the ceiling in sessions 1 or 2, the data of that subject were excluded from further analysis because the maximal performance of that patient could not be analyzed. Of the remaining patients, means, standard deviations, 95% confidence intervals, ICCs and, limits of agreement were calculated. An ICC of 0.75 or more was considered a measure for acceptable reliability (10–13). A kappa value of more than 0.60 was considered an acceptable reliability (14). Arbitrarily, a percentage of absolute agreement of 80% or more was also considered an acceptable reliability. All analyses were performed in SPSS.

RESULTS

Of the 30 subjects included, 27 subjects completed both sessions. Three patients did not attend session 2, stating that they did not feel capable of any manual handling due to LBP. Partial data sets were obtained from two subjects because of lack of time to complete testing, and these are reflected in the number of subjects per test in the tables.

Material-Handling Tests and Shuttle-Walk Test

The results of the test–retest reliability of the material-handling tests and shuttle-walk test yielded ICC values ranging from 0.75 to 0.87 (Table II). Limits of agreement ranged from 14.4 to 21.4 kg. Limits of agreement could not be calculated for the shuttle walk test because there was a systematic difference between the first and second sessions (9).

Criterion and Ceiling Tests

The results of the reliability of tests with a ceiling or a criterion are presented in Tables III and IV. Kappa values of 0.60 or higher were found for seven tests (Table III). For seven tests, kappa could not be calculated because of lack of filling of the cells in the 2×2 tables. Percentage of absolute agreement varied from 78 to 100% (Table III). The results of the additional analyses of the tests with a ceiling or a criterion are presented in Table IV. Of the six tests with a criterion, the ICC values ranged from 0.39 to 0.82. Only dynamic squatting reached the level of 0.75. Limits of agreement values ranged from 13.1 to 23.3 s. For the four tests with a ceiling, the ICC's ranged from 0.36 to 0.96 (Table IV). Only one test (forward bend test standing) reached the level of 0.75. The limits of agreement ranged from 63.4 to 102.2 s. For one item (the overhead work test), limits of agreement could not be calculated because of systematic differences between the first and second sessions. No further analyses were performed on the kneeling test, because only 11 subjects did not met the ceiling.

Summary of the Results

All 28 tests of the IWC FCE were divided into tests with and tests without an acceptable reliability on the basis of the kappa values, the percentage of absolute agreement and on the basis of the ICC values (Table V). Based on kappa values and percentage of absolute

Table II. Results of Paired *t*-Test, Limits of Agreement, and ICC's of the Material Handling Tests of the Modified Isernhagen Work System FCE and the Shuttle Walk Test

Activity (n paired observations)	Mean 1	SD 1	Mean 2	SD 2	Mean difference	SD	95%CI of difference	Limits of agreement	ICC	95% CI of ICC
Lifting in kg (27)	31.0	13.7	29.3	17.4	1.7	9.6	−2.1–5.5	±19.8	0.81	0.63 to 0.91
Overhead lifting in kg (27)	16.4	7.1	15.9	7.3	0.6	3.7	−0.9–2.1	±7.6	0.87	0.73 to 0.94
Short carry two-handed in kg (27)	36.6	16.1	35.2	17.6	1.4	10.4	−2.7–5.5	±21.4	0.81	0.63 to 0.91
Long carry two-handed in kg (27)	35.3	15.2	33.6	15.5	1.7	9.5	−2.1–5.5	±19.6	0.81	0.62 to 0.91
Long carry right-handed in kg (27)	27.3	10.8	27.6	11.7	−0.4	7.0	−3.2–2.4	±14.4	0.81	0.63 to 0.91
Long carry left-handed in kg (27)	27.6	10.7	26.5	11.7	1.0	7.0	−1.7–3.8	±14.4	0.81	0.63 to 0.91
Pushing static in kg (27)	38.7	11.4	40.7	10.2	−2.0	7.5	−4.9–1.0	±15.5	0.75	0.53 to 0.88
Pulling static in kg (27)	47.1	13.4	49.8	16.2	−2.6	9.7	−6.5–1.2	±19.9	0.78	0.58 to 0.89
Walking in meters (24)	367.1	129.6	398.3	143.5	−31.3	72.0	−61.7–.8	—	0.84	0.67 to 0.93

Note. Mean 1: group mean in the first session, mean 2: group mean in the second session, ICC: Intraclass correlation (one-way random model), 95%CI: 95% confidence interval —: limits of agreement cannot be calculated because there is a systematic difference between the first and the second sessions.

Table III. Criteria and Ceiling for Test Termination, Kappas, and Percentage of Similar Test Behavior, for Different Tests of the Modified Isernhagen Work System FCE

Tests (*n* paired observations)	Statistical level	Criterion* /Ceiling**	*n* Subjects reaching criterion/ceiling in session 1	*n* Subjects reaching criterion/ceiling in session 2	*K*	Similar test behavior
Pushing dynamic (26)	Dichotomous	Pushing a weighted cart over 20 m*	26	26	#	100% (26/26)
Pulling dynamic (26)	Dichotomous	Pulling a weighted cart over 20 m*	26	26	#	100% (26/26)
Overhead work test (27)	Continuous	15 min**	0	1	#	96% (26/27)
Forward bend test standing (27)	Continuous	15 min**	0	0	#	100% (27/27)
Forward bend test sitting (27)	Continuous	5 min**	5	4	0.60	89% (24/27)
Kneeling (27)	Continuous	5 min**	15	11	0.57	78% (21/27)
Crawling (25)	Dichotomous	10 repetitions*	23	23	1.00	100% (25/25)
Crouching (27)	Dichotomous	60 seconds*	23	24	0.84	96% (26/27)
Dynamic bending (27)	Continuous	20 repetitions*	20	21	0.70	89% (24/27)
Dynamic squatting (27)	Continuous	20 repetitions*	18	19	0.91	96% (26/27)
Rep. rotation standing right (27)	Continuous	30 repetitions*	23	21	0.51	85% (23/27)
Rep. rotation standing left (26)	Continuous	30 repetitions*	22	20	0.58	85% (23/27)
Rep. rotation sitting right (27)	Continuous	30 repetitions*	21	22	0.87	96% (25/26)
Rep. rotation sitting left (26)	Continuous	30 repetitions*	20	20	0.78	92% (24/26)
Stair climbing (27)	Dichotomous	20 x 5 steps up/down*	10	14	0.56	78% (21/27)
Ladder climbing (26)	Dichotomous	5 times up/down*	24	24	0.25	85% (23/27)
Balance (total of 6 tests) (25)	Dichotomous	less than 6 failures *	25	26	#	96% (25/26)
Sitting tolerance (26)	Dichotomous	30 min*	28	26	#	96% (25/26)
Standing tolerance (27)	Dichotomous	30 min*	26	26	#	93% (25/27)

Note. K: Kappa, #: Kappa values cannot be calculated due to lack of filling of the cells in the 2 × 2 table.

*Criterion indicates that a test is fulfilled if the criterion is met. For instance, the test of pushing dynamic has as criterion that a subject is able or unable to push a weighted cart over a distance of 20 m safely.

**Ceiling indicates that the test is terminated because the patient has met what is defined to be the maximal time of performance. For instance, working static overhead has a ceiling effect at 15 min. The test is terminated when the subject reaches 15 min. However, in that case the subject has not performed to his/her maximal ability.

Table IV. Results of Paired *t*-Test, Limits of Agreement, and ICC's for Tests (With a Criterion* or a Ceiling** Effect) of the Modified Isernhagen Work System FCE

Activity in seconds (*n* paired observations)	Mean 1	SD 1	Mean 2	SD 2	Mean difference	SD	95%CI	Limits of agreement	ICC	95%CI
Overhead work test** (26)	246.7	97.9	191.7	80.0	55.0	91.5	18.1 to 91.9	—	0.36	−0.02–0.65
Forward bend test standing** (27)	137.7	101.5	141.1	106.2	−3.3	30.9	−15.5 to 8.9	±63.4	0.96	0.91–0.98
Forward bend test sitting** (21)	131.7	68.6	132.7	59.7	−1.0	49.0	−23.3 to 21.4	±102.2	0.72	0.44–0.88
Dynamic bending* (19)	55.1	14.2	51.1	9.4	4.0	8.4	−0.01 to 8.1	±17.7	0.72	0.41–0.88
Dynamic squatting* (18)	50.3	11.3	48.3	9.8	2.0	6.2	−1.1 to 5.1	±13.1	0.82	0.59–0.93
Rep. rotation standing right* (20)	75.4	8.7	75.2	7.4	0.3	7.4	−3.2 to 3.7	±15.4	0.60	0.24–0.82
Rep. rotation standing left* (19)	73.1	7.5	71.2	6.4	1.8	7.7	−1.9 to 5.5	±16.2	0.39	−0.06–0.71
Rep. rotation sitting right* (21)	77.7	9.8	79.1	8.5	−1.4	7.8	−5.0 to 2.2	±16.4	0.64	0.30–0.83
Rep. rotation sitting left* (19)	75.3	9.5	77.0	11.5	−1.7	11.1	−7.1 to 3.6	±23.3	0.45	0.02–0.74

Note. Mean 1: group mean in the first session, Mean 2: group mean in the second session, ICC: Intraclass correlation (one-way random model), 95%CI 95% confidence interval, Limits of agreement cannot be calculated because there is a systematic difference between the first and the second session.

*Only if a subject reached the criterion in session 1 and session 2, were the times needed to reach the criterion in the sessions used for further analyses; other subjects were excluded from the analysis.

**If a subject reached the ceiling in session 1 or session 2, he/she was excluded from further analyses because the maximal performance cannot be analyzed.

Table V. Overview of the Reliability of the Items of the Modified IWS FCE

Items	Kappa, agreement %		Intra Class Correlation	
	Analysis	Reliability	Analysis	Reliability
Lifting	N/A	N/A	ICC high	acceptable
Overhead lift	N/A	N/A	ICC high	acceptable
Short carry two-handed	N/A	N/A	ICC high	acceptable
Long carry two-handed	N/A	N/A	ICC high	acceptable
Long carry right-handed	N/A	N/A	ICC high	acceptable
Long carry left-handed	N/A	N/A	ICC high	acceptable
Pushing static	N/A	N/A	ICC high	acceptable
Pulling static	N/A	N/A	ICC high	acceptable
Pushing dynamic	Kappa not calculated, agreement (%) high	acceptable	N/A	N/A
Pulling dynamic	Kappa not calculated, agreement (%) high	acceptable	N/A	N/A
Overhead work test	Kappa not calculated, agreement (%) high	acceptable	ICC low	not acceptable
Forward bend test standing	Kappa not calculated, agreement (%) high	acceptable	ICC high	not acceptable
Forward bend test sitting	Kappa high, agreement (%) high	acceptable	ICC low	acceptable
Kneeling	Kappa low, agreement (%) low	not acceptable	N/A	N/A
Crawling	Kappa high, agreement (%) high	acceptable	N/A	N/A
Crouching	Kappa high, agreement (%) high	acceptable	N/A	N/A
Dynamic bending	Kappa high, agreement (%) high	acceptable	ICC low	not acceptable
Dynamic squatting	Kappa high, agreement (%) high	acceptable	ICC high	acceptable
Rep. Rotation standing right	Kappa low, agreement (%) high	not acceptable	ICC low	not acceptable
Rep. Rotation standing left	Kappa low, agreement (%) high	not acceptable	ICC low	not acceptable
Rep. Rotation sitting right	Kappa high, agreement (%) high	acceptable	ICC low	not acceptable
Rep. Rotation sitting left	Kappa high, agreement (%) high	acceptable	ICC low	not acceptable
Walking	N/A	N/A	ICC high	acceptable
Stair climbing	Kappa low, agreement (%) low	not acceptable	N/A	N/A
Ladder climbing	Kappa low, agreement (%) high	acceptable	N/A	N/A
Balance	Kappa not calculated, agreement (%) high	acceptable	N/A	N/A
Sitting tolerance	Kappa not calculated, agreement (%) high	acceptable	N/A	N/A
Standing tolerance	Kappa not calculated, agreement (%) high	acceptable	N/A	N/A

Note. N/A: not applicable.

agreement, 15 of the 19 tests (79%) showed an acceptable agreement. Based on ICC values, 11 of the 18 tests (61%) showed an acceptable reliability.

DISCUSSION

Material-Handling Group and Shuttle-Walk Test

All eight tests of the material-handling group and the shuttle-walk test had ICC values above 0.75. This indicates that the variance in the test results between patients is considerably higher than the variance in test results within subjects. The ICC is a ratio between the signal (between-subject variance) and the signal plus noise (within-subject variance). Based on these ICC's, it can be concluded that these tests are reliable. However, the ICC

only expresses how well two observations are likely to classify a patient consistently relative to the other patients (15). The ICC value provides no indication of the magnitude of the disagreement between two observations (within patient variance: "noise") (16). To determine the magnitude of disagreement on an individual level, the limits of agreement were calculated (9). The limits of agreement of most of the material-handling tests were large. This means that the "noise" was relatively large, despite high ICC values. For example, for lifting an ICC value of 0.81 was found, but the limits of agreement were ±19.8 kg (mean performances 31.0 and 29.3 kg). In other words, approximately 95% of all differences within subjects will lie between ±19.8 kg.

Large limits are the result of a large within-patient variance. This variance can be attributed to the testing procedure, differences in interpretation of the evaluator, measurement errors and random error of the testing procedure, or to other factors such as the patient, differences in test behavior due to disparities in pain or motivation, or within-patient random errors. Without formally controlling for this, we hypothesize that a major part of the variance can be attributed to the patient (3). Because the limits of agreement for the IWS tests have not been investigated before, it is not possible to compare our results with those of other researchers. Therefore, we decided that, based on a statistical decision (ICC), the material tests and the shuttle walk test are reliable. However, a considerable amount of noise should be taken into account when interpreting the test results clinically.

Criterion and Ceiling Tests

For seven tests, kappa values and percent agreement were good (forward bend test sitting, crawling, dynamic bending, crouching, dynamic squatting, repetitive rotation sitting right and left). For seven other tests, kappa values could not be calculated due to lack of filling of the cells (pushing dynamic, pulling dynamic, overhead work test, working static standing forward, balance, sitting tolerance, and standing tolerance). However, their percentages of absolute agreement were very high (96–100%), and are therefore considered reliable as well. For three tests, kappa's were (far) below 0.60 and percentage of absolute agreement above 0.80 (repetitive rotation standing right, left, and ladder climbing), and for two tests (kneeling and stair-climbing) kappas as well as percentage of absolute agreement were below the criteria for acceptance. An explanation for this discrepancy is probably the lack of variation in cell fillings (13). In our study there is a large proportion of agreement, most of which is limited to only one of the possible rating choices. Under this limited variation, only one decision can make the difference between poor and excellent reliability expressed as a kappa. In our study this phenomenon has resulted in some low kappa values and (very) high percentage of absolute agreement. The use of percentage of absolute agreement has its limitations as well, because it does not take into account the agreement that is expected to occur due to chance alone (16). Cohen's kappa, on the other hand, corrects the observed agreement for the agreement that is expected by chance. Because both measures of reliability contain their strength and their limitations, it was decided to apply both in this study.

Before analysis, patients who reached the ceiling in sessions 1 and 2 were excluded for further analyses because the maximal performance could not be analyzed. If a patient reached the criterion in sessions 1 and 2, the time needed to reach the criterion was used for further analyses. Those subjects who did not reach the criterion were excluded. Only two tests (forward bend test standing and dynamic squatting) reached the ICC level of

0.75. Two tests show ICC values of 0.72 (forward bend test sitting and dynamic bending). Therefore, the reliability of these tests is disputable. For six tests, ICC values were low. The limits of agreement (Table IV) are relatively smaller compared to those of material handling tests (13.1–17.7 s). Of the tests with ceiling or criterion effects, only one test (dynamic squatting) shows a high ICC and small limits of agreement. Based on a statistical decision (ICC), only forward bend test standing and dynamic squatting are reliable. For static standing, a large amount of within-patient variation ("noise") should be taken into account when used clinically.

Similar relationships were found between kappas or percentage agreement (high values) and ICC's (low values) for the overhead work test, repetitive rotation sitting right and left. This indicates that a high percentage of similar test behavior (reaching a ceiling or a criterion, or not) does not predict high ICC's (quantitatively the same test behavior in the two sessions). These measures for reliability describe clearly different aspects of reliability.

The tests of the IWS FCE were acceptable to all subjects and required no specialized equipment. Some patients reported an increase in pain while performing tasks, others reported it 1 or 2 days after the testing day. Two weeks separated the first and second testing sessions. In this time period no significant change in work ability was expected, yet time was allowed to lessen recall of the previous test results and to recuperate from the first test session (12). Despite the use of the same evaluator for both test sessions, significance differences occurred between both sessions for the shuttle-walk test and overhead work test. This indicates that the first session may have influenced the results of the second session. It can be debated whether this influence is a form of physiological training or that knowledge of the first test session by the patients influences their test behavior in the second test session. Furthermore, these significant differences may have occurred by chance due to multiple statistical testing. For the other tests, no significant difference occurred. Conflicting results were found for the repetitive rotation tests. For rotation sitting right and left high kappa values were found (0.87 and 0.78, respectively) while for standing right and left lower kappa values were found (0.51 and 0.58, respectively). These differences in kappa values cannot be explained satisfactorily.

Selection bias may have influenced our test–retest results. In 1 year, out of approximately 100 patients who met the selection criteria, only 30 were willing to participate. Main reasons for not participating were that testing would take too much time. As a result, only those subjects who were motivated and had time participated in our study. This means that our study sample, and therefore also test–retest reliability, may differ from the population in clinical practice. In this study, protocols were used that were slightly modified from the original IWS FCE. Reliability of the original IWS FCE tests should be analyzed in future research. Basically, the IWS FCE is a set of tests with very heterogeneous properties, ceiling and criterion tests. In analyzing the reliability, different types of analysis had to be performed and sometimes arbitrary decisions had to be taken. This resulted in rather complex results which could not be interpreted simply. Changes in the testing procedure, for example to eliminate ceiling-effects, may improve reliability.

CONCLUSION

Test–retest reliability of 15 tests (79%) of the modified IWS FCE was acceptable based on kappa values and percentage of absolute agreement. For 11 tests (61%), test–retest reliability was acceptable based on ICC values.

ACKNOWLEDGEMENT

The authors thank Rita Schiphorst-Preuper, Cor Muskee, Willem Jorritsma, and Janine Stubbe for their valuable assistance in selecting patients and collecting the data. This study was supported by Zorgonderzoek Nederland (ZON), Grant Number 96-06-006.

REFERENCES

1. King PM, Tuckwell N, Barrett TE. A critical review of Functional Capacity Evaluations. *Phys Ther* 1998; 78: 852–866.
2. Abdel-Moty E, Fishbain DA, Khalil TM, Sadek S, Cutler R, Rosomoff RS, Rosomoff HL. Functional capacity and residual functional capacity and their utility in measuring work capacity. *Clin J Pain* 1993; 9: 2003–2013.
3. Gross DP, Battie MC. Reliability of safe maximum lifting determinations of a functional capacity evaluation. *Phys Ther* 2002; 82: 364–371.
4. Reneman MF, Dijkstra PU, Westmaas M, Göeken LNH. Test-retest reliability of lifting and carrying in a 2-day functional capacity evaluation. *J Occup Rehab* 2002; 12: 269–275.
5. Hart DL. Test–retest reliability of the static push/pull tests for functional capacity evaluations. *Phys Ther* 1988; 68: 824.
6. Reneman MF, Bults MMWE, Engbers LH, Mulders KKG, Goeken LNH. Measuring maximum holding times and perception of static elevated work and forward bending in healthy young adults. *J Occup Rehab* 2001, 11(2): 87–97.
7. Isernhagen Work Systems. *Functional Capacity procedure manual* 1st edn. Duluth, MN, 1997.
8. Reneman M.F., Jaegers SMHJ, Westmaas M, Göeken LNH. The reliability of determining effort level of lifting and carrying. *Work* 2002; 18: 23–27.
9. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; 8: 307–310.
10. Lee J, Koh D, Ong CN. Statistical evaluation of agreement between two methods for measuring quantitative variable. *Comput Biol Med* 1989; 19: 61–70.
11. (a) Tammemagi MC, Frank, JW, LeBlanc M, Artsob H, Streiner DL. Methodological issues in assessing reproducibility—A comparative study of various indices of reproducibility applied to repeat elisa serologic tests for lyme disease. *J Clin Epidemiol* 1995; 9: 1123–1132. (b) Altman DG. *Practical statistics for medical research.* London: Chapman & Hall, 1991.
12. Streiner DL, Norman GR. *Health measurement scales. A practical guide to their development and use*. 2nd edn. Oxford: Oxford University Press, 1995.
13. Innes E, Straker L. Reliability of work-related assessments. *Work* 1999; 13: 107–124.
14. Altman DG. *Practical statistics for medical research.* London: Chapman & Hall, London, 1991, p 404
15. Evans WJ, Cayten CG, Green PA. Determining the generalizability of rating scales in clinical settings. *Medical Care* 1981; XIX: 1211–1220.
16. Haas M. Statistical methodology for reliability studies. *J Manipulative Physiol Ther* 1991; 14: 119–132.